Which Academic Subjects have Most Online Impact? A Pilot Study and a New Classification Process

Mike Thelwall[1]
School of Computing and IT, University of Wolverhampton, 35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK. E-mail: m.thelwall@wlv.ac.uk
Liwen Vaughan
Faculty of Information and Media Studies, University of Western Ontario, London, Ontario, N6A 5B7, Canada. E-mail: lvaughan@uwo.ca
Viv Cothey
School of Computing and IT, University of Wolverhampton, 35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK. E-mail: viv.cothey@wlv.ac.uk
Xuemei Li
School of Computing and IT, University of Wolverhampton, 35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK. E-mail: x.li4@wlv.ac.uk
Alastair G. Smith
School of Information Management, Victoria University of Wellington, New Zealand.
Email: Alastair.Smith@vuw.ac.nz

Abstract

It is known that use of the Web by academic researchers is discipline-dependent and highly variable. Moreover, the Web is increasingly central for sharing information, disseminating results and publicizing research projects and so it is important to find out as much as possible about how it is used, and who is using it successfully. In this pilot study we seek to identify the subjects that have the most impact on the Web, and look for national differences in online subject visibility. We use link counts to identify the highest impact Web sites from the universities in Taiwan and Australia and classify them by domain type and by subject content. The highest impact sites were from computing, but there were major national differences in the impact of engineering and technology sites. Another difference was that Taiwan had more high impact non-academic sites hosted by universities. As a pilot study, the classification process itself was also investigated and the problems of applying subject classification to academic Web sites discussed. The study draws out a number of issues in this regard that do not have simple solutions and point to the need to interpret the results with caution.

Introduction

The Web is now firmly embedded into the fabric of universities, at least in the richer countries, and is being increasingly used in conjunction with research: to share information; to disseminate findings; to publicize projects, teams and departments. Given this extensive use and the resources now spent on Internet access and maintaining a Web presence, there is clearly a need to evaluate the effectiveness of the use that universities, departments and individual scholars make of the Web. This is a difficult task because of the hybrid nature of the typical university Web site (Middleton et al., 1999) and the variation in the extent to which Web use is natural to a discipline or field (Kling & McKim, 1999, 2000). It would also be helpful to identify the disciplines that are making the most effective use of the Web, and to find out why. This would shed some light on the use of this key communication medium and could lead to more effective academic use of the Web. Policy makers and research managers can use this information to help identify areas of best practice that can be used as a model for others. Increasingly also, evaluations of university and research sites take account of their Web presence. But are all disciplines and countries represented equally on the Web? If some disciplines inherently have more impact on the Web, others could be disadvantaged in research evaluation exercises.

In addition to research into scholarly use of the Web, more general studies of the people and organisations that publish on the Web are also possible. From the latter perspective, the Web is a new easily

[1] Thelwall, M., Vaughan, L., Cothey, V., Li, X. & Smith, A. (2003). Which academic subjects have most online impact? A pilot study and a new classification process, *Online Information Review* 27(5), 333-343.

accessed resource through which researchers can get unparalleled access to basic but timely information about the developed world (Weare & Lin, 2000), for instance information about the commonalities and differences between higher education systems internationally. Information scientists now have the opportunity to be at the forefront of developing effective and robust methodologies to exploit this rich information source.

Toward this end, we conduct a pilot study that uses Web hyperlink data to identify the highest impact academic Web sites for two different countries (Taiwan and Australia), categorizing the sites by domain type and subject content in order to uncover some characteristics of the two national education systems. We also discuss the effectiveness of this new methodology.

Literature Review

International comparisons of academic systems have been made possible with bibliometric analyses of databases of scientific publications, principally those of the Institute for Scientific Information (ISI). An example is a study of international co-authorship patterns in three disciplines, showing clear ties between particular pairs of nations (Glänzel & Schubert, 2001). This type of study tracks patterns of formal scholarly communication whereas Web based analysis is likely to involve a wide range of types of informal communication (Wilkinson et al., 2003). The advantage of the bibliometric approach is its methodological soundness and the years of theoretical background studies that make its results relatively well understood (Borgman & Furner, 2002) albeit controversial (Moed, 2002). The advantage of Web-based analysis, called Webometrics or cybermetrics, is unrestricted access to the data (the public Web) and its relative timeliness. The specific Webometric questions addressed here concern disciplinary and international differences in online impact. This is really a social issue concerning the way in which scholars are making use of the new technology.

The social aspects of computer technology are the concern of social informatics. From this field it was predicted that Web use would continue to vary by discipline and field because of differing communication needs (Kling & McKim, 2000). For example in some scientific fields there is a need for extensive collaboration and data sharing, often between remote scientists which can be facilitated in the form of a Web 'collaboratory' (Finholt, 2002). There are also vastly differing patterns of formal and informal communication in the offline world. For instance, journal articles in the hard sciences are typically the dominant medium but some also give importance to published "letters", whereas in the humanities books have a much greater importance (Hyland, 2000). There also appears to be a much greater degree of overall informal communication among humanities scholars, which would seem to be a natural reason for them to use the Web more. However, the opposite appears to be the case (Tang & Thelwall, 2003), possibly due to less access to and mastery of the technology relative to researchers in other fields.

So which methods can be used to assess online impact? Many researchers have suggested that hyperlinks possess many attributes of the citations that are often used to measure the impact of journals and articles (Larson, 1996; Rodriguez Gairín, 1997; Rousseau, 1997; Almind & Ingwersen, 1997; Ingwersen, 1998; Davenport & Cronin, 2000; Cronin, 2001) including Borgman and Furner (2002) in their recent review of bibliometrics. In fact, given the invalidity of Web site hit counters, Web links appear to be the only publicly accessible entities that are potential candidates for use in impact measures. Many studies have now used them, primarily to assess their validity, and recent results indicate that they can be a useful measure. This has been achieved primarily by demonstrating significant correlations with traditional measures of scholarly productivity (Thelwall, 2001c; Smith & Thelwall, 2002; Thelwall & Tang, 2003). The 'impact' that counts of links to an academic Web site seems to measure is an aggregation of a variety of types of informal scholarly communication (Wilkinson et al., 2003).

The usefulness of links as impact indicators is supported by computer science research into search engine design. The key Information Retrieval (IR) task for modern Web search engines is to find Web pages or sites that are likely to satisfy a user's information needs, based upon just a few query words submitted. Early search engines used simple text matching techniques. However, as the Web grew in size, the retrieval problem was transformed from one of finding a page containing the words to one of filtering out the matching pages that were unlikely to be particularly relevant. A natural way to do this was to differentiate between the more and less important parts of the page as well as counting the relative frequency of the matching words in a Web page. Google broke the mould when its founders, Brin and Page

(1998) based its crawling and ranking algorithms upon the link structure of the Web. The basic idea was that highly linked to pages were more likely to contain useful information than ones that nobody thought it worth linking to (Kleinberg, 1999). Google's commercial success is a very powerful argument for this, although there is little scientific evidence that the approach works (e.g. Hawking et al., 2000).

Research Questions

The issue of which subjects have the most impact on the Web will be addressed by formulating a concrete answerable question based upon a series of simplifications. First, it will be assumed that it is reasonable to equate the (rather fluid) term Web site with a collection of pages sharing a common domain name. Therefore a university Web space in our terminology would typically be a large number of Web sites, most having a domain name ending in the official root domain name of the university. So, for example, all Web pages with URLs starting with www.ucla.edu would be counted as one document, but those starting with www.gseis.ucla.edu would be counted as a separate one. Note that both end in the root domain name of the parent university, ucla.edu. Second, Web site impact will be estimated by the number of links to a site, its inlink count, from other universities in the same country. The assumption is that a high impact site will have many inlinks. Third, it will be assumed that it is reasonable to assess the online impact of a subject within a country by measuring the online impact of the Web sites for the subject. The specific questions addressed are as follows.

- Which subjects have the highest impact on the Web?
- Are there significant national differences in the highest impact subjects?
- Is it possible to construct a classification scheme for the subject content of Web sites that is relatively unambiguous in terms of being capable of delivering a high degree (e.g. >80%) of inter-classifier agreement?

The Classification Schemes

The Domain Type Classification Scheme

The following site types were recognized in the scheme.

1. University home domain
2. Library domain (physical, not virtual or digital library)
3. Bulletin board
4. E-journal domain or digital library of externally created documents of any kind
5. Resource (external - not created by staff or students of the hosting university, e.g. Sunsite, Tucows, mirror sites)
6. Subject-based internally created site (e.g. department/faculty sites, and including portal sites where links to other sites are included, but not the content of those sites)
7. Other, including other non-subject based sites

All sites classified as subject-based (number 6) were then further classified using the subject classification scheme below.

The Subject Classification Scheme

An international coding scheme from UNESCO (http://www.usc.es/citt/privado/codigos/unesco.htm) was chosen for the purpose of subject classification. The standard Library of Congress and Dewey Decimal schemes were not used in order to avoid the possibility of national or linguistic bias. The UNESCO coding scheme does not seem to be used extensively by the either of the two countries in our sample (although it is in Spain), and so this should not be a source of bias. The UNESCO scheme is highly detailed and previous research has shown the difficulty in assigning categories even to single pages (McMillan, 2000; Weare & Lin, 2000; Wilkinson *et al*., 2003) and so we decided to employ a coarse categorization approach using only the top-level categories. We made one modification to make computer science a top-level category. Computer science is heavily represented on the Web but it only appears as a subcategory of maths in the UNESCO scheme. We also merged a number of similar categories that we believed would be difficult to

separate in practice. Table 1 in Appendix 2 shows the final classification scheme used and the counterpart categories in the UNESCO scheme.

Another alternative scheme that was investigated but rejected was the Open Directory Project's directory structure (dmoz.org) – or at least the academic portion of it. This is created by hundreds of thousands of volunteers and has the advantage of being fast and updateable through being online. Although the structure is far from dominated by academic content, it is represented and could possibly have formed the basis for a scheme. For example, the major academic areas can be found. Art is at dmoz.org/Arts/, Science at dmoz.org/Science/, Humanities at dmoz.org/Arts/Humanities/, and Social Sciences at dmoz.org/Science/Social_Sciences/. This scheme is not suitable since its categorisers are probably from the US in the majority, and perhaps biased towards computer science. Unsurprisingly, there is a whole category for computing dmoz.org/Computers/. Some academic departments are in the directories and classified, but again we did not want to rely upon the judgements of an unknown group of people.

Finally, we should mention that there have been attempts to automatically classify Web documents, e.g., through the Dewey Decimal scheme (Jenkins *et al*., 1998) and others (Devadason *et al*., 2001). The attractiveness of automatic indexing would be scalability and reproducibility but it would still be language-dependant and is much too unreliable to be seriously considered.

Methods

Link Data Collection and Processing

We chose Australia and Taiwan as the two countries to compare in the study because of the availability of data about their link structure, and as examples of developed countries that we suspected could have different patterns of Web use due to differing linguistic and cultural backgrounds. Developed countries were chosen because developing countries' universities may not have enough of a Web presence for the purposes of the study. The link structures of Australian universities (collected from October, 2001 to January, 2002) and Taiwanese universities (collected in February to March, 2002) were obtained from the free online database at cybermetrics.wlv.ac.uk/database. This is believed to be the only source of such data and so although ideally a random selection of countries would be used for this research, in practice this is not possible. Only university sites are included in the data set, excluding all other types of academic site. Descriptions of the data source and collection process have been published (Thelwall, 2001a, 2001b).

Counting links between pages directly for developing visibility indicators is known to be problematic for two reasons. First, internal site links typically dominate others, but offer a relatively low quality indicator of the value of the target page, and this argument extends to links between different Web sites belonging to the same university (Thelwall, 2000). As a result, only links between different universities (in the same country) will be considered. Second, links between sites can be automatically generated in huge numbers, for example in the navigation bar of a joint project Web site (Thelwall, 2002a). We will use a method that has been developed to circumvent this and other problems, to count links between (our domain-based) Web sites instead of between pages (Thelwall, 2002a). In other words, a link will be from one Web site to another when the Web sites come from different universities and at least one page from the first contains a link to at least one page of the second. Subsequent links between the same (ordered) pair of Web sites are ignored; the maximum count is 1. For example, the link count from Web site www.ucla.edu to www.drexel.edu would be either zero or 1. If any page with URL starting www.ucla.edu (as its domain name) contained a link to any page with URL starting www.drexel.edu then the count would be 1, otherwise it would be zero. In previous research this is known as the domain alternative document model counting method.

This approach makes the counting process dependant on the internal organization of Web sites. In particular, universities that use a lot of different domain names would have more 'Web sites' using our definition, and therefore inflated link counts. Despite this reservation, link counting approaches using (domain-based) Web sites have been shown to be statistically more reliable than other counting methods such as those based on individual pages, at least for the UK universities (Thelwall, 2002a).

The Classification Process

Each Web site was classified first by matching according to the appropriate non-subject scheme described above first, and those matching type 6, subject-based internally created sites, were allocated a

subject from Table 1 of Appendix 2. In cases where more than one subject appeared to have a significant presence in the Web site, all the subjects were recorded. The judgment of subject content was made based upon the home page description. If the subject content was not clear from the home page, other pages in the site were visited. If a site was no longer available, the Internet Archive WayBack Machine (www.archive.org) was used to retrieve and view the Web pages for classification purposes. The Archive does have national biases in coverage (Thelwall & Vaughan, 2004), but the number of pages tested in it should not have been large enough to influence the results.

The classification process was as follows. Each country was allocated to two independent native language speakers for classification. A sorted list of Web sites, from the highest linked to lowest linked, was provided to each classifier. The classification of sites on the list continued until 100 subject-based sites had been categorized. The total numbers of sites classified (including those not subject-based) were 194 and 209 for Australia and Taiwan respectively. Comments in English were recorded for each site to justify the categorization. The two classification results for each country were then compared and a single researcher used the comments to adjudicate on differences and attempt to ensure consistency of categorization across the different countries. Domains with multiple classifications were assigned fractional weights for the purpose of collating results. For example if a site was classified into three different categories, then a weight of 1/3 was assigned to each of the three categories.

An example of an easy classification decision was http://www.edfac.usyd.edu.au, a page with the title "Welcome to the Faculty of Education", and classified as 16, Education (see Table 1). A more difficult one was http://www.nisu.flinders.edu.au, the home page of the Research Centre for Injury Studies. One classifier chose 20, sociology for this and the second chose 8, medical sciences because the site contained a mixture of epidemiology and social concerns. Following links from the home page would help the classifier to see the kind of personnel involved with the centre, its funding source and the kind of information available, all of which would help to identify it as a public health site, which could be fitted in to the UNESCO classification scheme (medical science, public health, No. 3232) after investigating it at a more detailed level than shown in Table 1.

Results and Discussion

Inter-classifier Consistency

Figures 1-3 give the results of the classifications and an analysis of the extent of inter-classifier agreement. In just over 60% of cases both classifiers agreed exactly on the subject classification and in just under 10% they disagreed completely. In the majority of the remaining disagreements, one classifier assigned more categories than the other (extension). There was little difference by country in inter-classifier consistency. The larger percentage for Taiwan in the 'extension' class could be explained by the fact that one classifier categorized many computer science departments as 'Technology and Engineering' in addition to 'Computing', if they had evolved from a physics/electronics background rather than a maths background. This was a reasonable assumption, but one that our classification scheme was not sophisticated enough to confirm or deny.

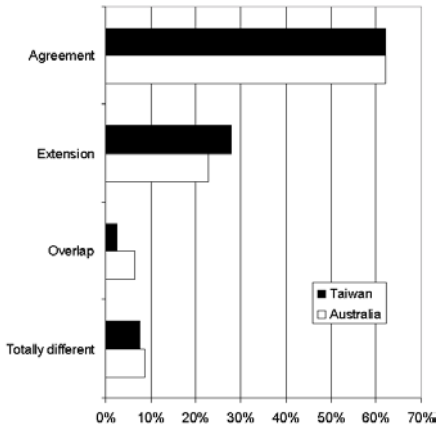

Figure 1: The types of agreement between classifiers on the subject contents of academic Web sites.

Web Site Type Comparisons

Figure 2 shows the broad similarity between the countries in Web site types. However, Taiwan had more 'other' category sites and less subject-based sites. The 'other' category included non-academic sites such as what appeared to be an online youth arts magazine (http://r703a.chem.nthu.edu.tw), with the description meta tag (in English): "Anthologies and E-Zines created by dirty underground artists in Taiwan, including comix, photography, computer graphics, etc". We investigated this example further to see why it had attracted so many academic links. In fact most of them pointed to just one page on the site, an instruction manual in Chinese on how to create Web pages (http://r703a.chem.nthu.edu.tw/handbook/handbook.html). This page was clearly an integral part of the Web site (rather than a completely independent publication that just happened to be hosted on the same domain) because it contained the logo from the home page. This disturbingly indicates how the online impact of a site may not reflect its ostensible purpose. The variety of types of use of Taiwanese educational Web sites seems to be much greater than that of Australia, and perhaps the West in general (from the researchers' experiences in previous academic Web surveys). This seems to be a significant cultural/attitude difference.

Note that the differences between countries in Figure 2 are not statistically significant (using a Chi-squared test, merging the three small categories into the "other" category in order to satisfy the minimum cell count requirements). The above comparison between the two countries is therefore anecdotal in nature.

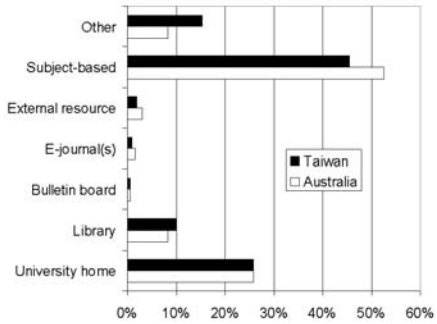

Figure 2: A breakdown of the identified types of high impact university sites in Australia and Taiwan.

Subject Content Comparisons

Figure 3 shows differences over almost the whole spectrum of subjects. The significance of small differences should not be exaggerated, however, since the numbers are relatively small and different classifiers obtained the results. As expected, computing features heavily, but the rest of the sciences are not as dominant as previous research would suggest (Thelwall, 2002b, Tang & Thelwall, 2003). In terms of differences, the 'Technology and Engineering' class is prominent. Taiwan appears to have an enormous number of high impact Web sites for this subject while Australia has almost none. This could reflect (a) a real difference in the national sizes of the disciplines, (b) a real difference in the national online impact of the disciplines, or (c) a classification difference due to classifier perceptions or the self-description of the Web sites. We revisited some of the sites and concluded that (a) was the most likely explanation. The 'Technology and Engineering' class of sites included a wide range of types of engineering and was dominated by electronic engineering. One of the sites was a department of 'Computer Science and Information Engineering', which was dual classified with Computing, but the rest were clearly classified correctly. Taiwan appears to have a very high specialism in electronic engineering in particular. From visiting the sites it is clear that the proliferation of this subject is a genuine phenomenon, and that it is not a naming ambiguity. Here is the description of one such site (Fu Jen Catholic University, 2003), which is electronics connected to computing, but clearly should not be classified as computer science.

> *The Department of Electronic Engineering at Fu Jen Catholic University was founded in 1977 to meet the country's need for economic growth and high tech development. Its purpose is to cultivate highly qualified professionals in electronic engineering, particularly in the areas of communications, control, computer, and very large scale integrated circuit design/computer aided design (VLSI/CAD).*

There is an interesting interpretation of this finding from a sociological perspective. A popular science studies theory has postulated that there is a new trend in science, dubbed 'Mode 2 Science' which involves, amongst other things, an increased use of communications technologies and closer cooperation between universities and industry (Gibbons et al., 1994; Kraak, 2000). The implication of this theory for Australia

has attracted some attention (Ronayne, 1997). The very high impact of technology and engineering in Taiwan is an indicator of a very healthy Mode 2 science sector in this country (although engineering does not have to be Mode 2). We speculate that the relatively newly expanded university sector is in this respect more modern than that of Australia.

Other categories with sizeable discrepancies that seem to be significant are Law, Medical Sciences and Life Sciences. Overall, there also seems to be more emphasis on the social sciences in the Australian academic Web, with the exception of economics.

A Chi-squared test was used to assess whether the differences in the graph were statistically significant. We merged the categories into four: Science; Social Sciences and Humanities; Computing; Technology and Engineering. The national differences were found to be significant at the 0.01% level. The conclusion is that Taiwan has more high impact Web sites classified as Computing and Technology and Engineering while Australia has more classified as Science, and Social Science & Humanities. Note that this finding applies to the data collected, and must be interpreted in conjunction with the discussions over the reliability of methodology.

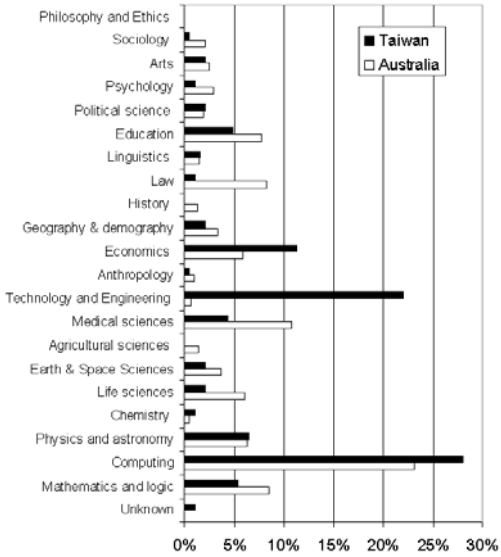

Figure 3: A subject breakdown of subject-based high impact university sites in Australia and Taiwan.

Discussion of Methodology

We will center the discussion upon issues that arose in the classification process because the Web site and link counting problems have already been addressed in the methodology. We will then make suggestions for improvements in the methodology.

Classifying Whole Sites

Any classification process must start from a clear statement of its purpose. In our case the classification scheme is not being prepared for an end user who will have an information need to satisfy, but for identifying the subjects that have a high online Web impact as measured by links. The link count reliability issues suggested that counting by (domain name based) Web site was better than counting by page, leading to the need to classify a number of academic sites that contained tens of thousands of pages. Our approach was to attempt to predict the content of the site by viewing its home page and perhaps a few others. The (untested) underlying assumption was that home page descriptions would broadly match site contents. In at least two cases this was not true: the home page was a default page giving information about the Web server software, but the server itself hosted a range of academic information. This was discovered by using an AltaVista site search. In a few other cases the home page was insufficiently informative so as to be misleading for our purposes. For example some Business schools were discovered to include Law in their portfolios, but it was not mentioned anywhere in their home page. To improve the reliability of the results two steps are needed. (1) Randomly sample the Web pages in a selection of Web sites and compare their content with the classification given to the site as a whole to see the proportion of sites that can be

accurately classified in this way. (2) Develop a protocol to ensure that sites are sufficiently investigated to give a high chance of accurate classification. This may mean drilling down to at least departmental level in all faculty and school sites, for example.

The Classification Scheme

We attempted to use an internationally neutral scheme, but it was not suited to the task because it was outdated and not designed for classifying tertiary education departments. To give some examples of this: geography and demography appear to be shrinking and being partly replaced by environmental science; computing has significantly shifted away from mathematics and has incarnations that are separately married with business, psychology and linguistics; electrical engineering can be packaged as applied maths, computing, or a type of engineering. Arts is another widely used university faculty designation, but does not correspond to a single body of published documents, and so is difficult to place in a document-based classification.

In retrospect, given the difficulties that we encountered, an alternative classification scheme from UNESCO may well have been easier to implement, its International Standard Classification of Education (UNESCO, 1997), which includes categories for the types of sites that were problematic to classify under the existing scheme and was more modern in its coverage of subjects.

An alternative classification approach that would be fully up-to-date is clustering rather than indexing: simply grouping together similar sites. This approach also has disadvantages: the need for eventual indexing to cope with multi-disciplinary sites and the difficulty in creating an international scheme that is fully comparable.

Alternative Methods

Given the evident difficulties in effectively classifying academic Web sites by subject, it is also worth investigating a different methodology to address the main question of the paper. As an example it may be worth investigating whether it is possible to develop a procedure to minimize the anomalies in links between individual pages so that pages rather than sites could be classified. Unfortunately, this would not make the exercise easier since links to site home pages of other universities are very common (Thelwall, 2002b) and these would presumably need to be classified on the basis of the content of the site as a whole anyway. Also, individual pages would probably be much harder to classify since at least site home pages typically give some form of accessible self-description aimed at outsiders, even if it is only the department/faculty name. It is necessary, however, to resolve the problem of sites having high online impact due only to content unrelated to their main purpose, as in the example above.

A methodological issue is that the classification process does not take into account factors other than subject content that can influence the creation of links, such as site usability and age. The Web is an evolving system and the snapshot used is likely to capture a combination of old and new pages. Inevitably there is a time delay between the creation of a page and other authors linking to it, and so newer sites may tend to have a higher actual impact than is reflected by their inlink counts. Methods have been devised to track time factors on the Web (Vaughan & Thelwall, 2003), which would provide an additional dimension to a future study.

Conclusions

The classification process overall found as expected that computing was heavily represented on the Web, but that the rest of science was less dominant over social sciences and humanities than previous studies would have suggested.

In terms of national differences, the clearest finding from the study was the relatively high online impact within Taiwan of technology and engineering, particularly electronic engineering. We speculate that this was evidence of a very healthy modernity in terms of the academic approach to science in Taiwan. The Mode 2 Science theory (Gibbons et al., 1994) is also a partial explanation for the difficulty in classifying academic Web domains. The theory includes the contention that the new mode of science, and in fact wider academic practice, is much more interdisciplinary (Kraak, 2000; Gibbons, 2000). One consequence of this would be the frequent coexistence of multiple different subjects in one Web site. It is natural that new titles are developed to describe the multidisciplinary research areas e.g. ‘computational linguistics’, ‘data

mining', 'cognitive science' and that outsiders will be unaware of the exact meaning of the descriptions, or perhaps even that the apparent meaning is different from the actual meaning to subject experts.

Those involved in a subject in one country that appears to have more online impact in the other are advised to investigate whether they are missing an opportunity. Similarly, those in low impact subjects are advised to assess whether they are missing out on the opportunities of the Web. Of course, other explanations are possible in individual cases, including different subject sizes, and so further investigations are needed to discover whether this is the case. Governments may wish to finance a full-scale version of the kind of Web based fact-finding exercise reported here in order to check the health of their academic Web. Financial backing would allow a much more extensive and finer grained classification exercise that could in conjunction with subject size information, e.g. the number of academics in each subject area, give reliable data from which to formulate policy and best practice guidelines with regard to Web use for research.

The information about the national education systems obtained from the study attests the value of this kind of research. However, the method itself was found to be problematic and in need of additional study. Inter-classifier consistency needs to be improved to ensure the validity and reliability of the data and a new classification scheme needs to be developed to cope with the relatively fast changing nature of academic subjects and their organization.

Acknowledgement

The first and third authors were supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technological Development of the European Commission. It is part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015). The referees are thanked for their helpful suggestions.

References

Aguillo, I. F. (1998). "STM information on the Web and the development of new Internet R&D databases and indicators". In: Online Information 98: Proceedings. Learned Information, pp. 239-243.

Almind, T. C. and Ingwersen, P. (1997). "Informetric analyses on the world wide Web: methodological approaches to 'Webometrics'." Journal of Documentation, Vol. 53 No. 4, pp. 404-426.

Borgman, C and Furner, J. (2002). "Scholarly communication and bibliometrics." In: Cronin, B. (ed.), Annual Review of Information Science and Technology 36, pp. 3-72, Medford, NJ: Information Today Inc.

Brin, S. and Page, L. (1998). "The Anatomy of a large scale hypertextual Web search engine." Computer Networks and ISDN Systems, Vol. 30 Nos. 1-7, pp. 107-117.

Cronin, B. (2001). "Bibliometrics and beyond: Some thoughts on Web-based citation analysis." Journal of Information Science, Vol. 27 No. 1, 1-7.

Davenport, E. and Cronin, B. (2000). "The citation network as a prototype for representing trust in virtual environments." In: Cronin, B. and Atkins, H. B. (eds.). The Web of knowledge: a festschrift in honor of Eugene Garfield. Metford, NJ: Information Today Inc. ASIS Monograph Series, pp. 517-534.

Devadason, F., Intaraksa, N., Patamawongjariya, P. and Desai, K. (2001). "Search interface design using faceted indexing for Web resources." Proc. ASIST 2001, pp. 224-238.

Finholt, T. (2002). "Collaboratories." In: Cronin, B. (ed.), Annual Review of Information Science and Technology 36, Medford, NJ: Information Today Inc., pp.73-107.

Fu Jen Catholic University (2003). Retrieved 17 January 2003 from http://www.ee.fju.edu.tw/English%20Instruction.html

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., and Trow, M. (1994). The New Production of Knowledge. London: Sage.

Gibbons, M. (2000). "Universities and the new production of knowledge: Some policy implications for government." In: Kraak, A. (ed.). Changing modes: New knowledge production and its implications for higher education in South Africa. Pretoria: HSRC Publishers, pp. 34-44.

Glänzel, W. and Schubert, A. (2001). "Double effort = double impact? A critical view at international co-authorship in chemistry." Scientometrics, 50(2), 199-214.

Hawking, D., Bailey, P. and Craswell, N. (2000). "ACSys TREC-8 experiments." In: Information Technology: Eighth Text REtrieval Conference (TREC-8), NIST, Gaithersburg, MD, USA, pp. 307-315.

Hyland, K. (2000), "Disciplinary discourses: social interactions in academic writing", Harlow: Longman.

Ingwersen, P. (1998). "The calculation of Web Impact Factors." Journal of Documentation, 54(2), 236-243.

Jenkins, C., Jackson, M., Burden, P., and Wallis, J. (1998). "Automatic classification of Web resources using Java and Dewey Decimal classifications." Computer Networks and ISDN Systems, Vol. 30 pp. 646-648.

Kleinberg, J. (1999). "Authoritative sources in a hyperlinked environment." Journal of the ACM, Vol. 46 No. 5, pp. 604-632.

Kling, R. and McKim, G. (1999). "Scholarly communication and the continuum of electronic publishing." Journal of the American Society for Information Science, Vol. 50 No. 10, pp. 890-906.

Kling, R. and McKim, G. (2000). "Not just a matter of time: field differences in the shaping of electronic media in supporting scientific communication." Journal of the American Society for Information Science, Vol. 51 No. 14, 1306-1320.

Kraak, A. (2000). "Changing modes: A brief overview of the mode 2 debate and its impact on South African policy formulation." In: Kraak, A. (ed.). Changing modes: New knowledge production and its implications for higher education in South Africa. Pretoria: HSRC Publishers, pp. 34-44.

Larson, R. R. (1996). "Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace." In: Proceedings of the AISS 59th annual meeting.

McMillan, S. (2000). "The microscope and the moving target: The challenge of applying content analysis to the world wide Web." Journalism and Mass Communication Quarterly, Vol. 77 No. 1, pp. 80-98.

Middleton, I., McConnell, M. and Davidson, G. (1999). "Presenting a model for the structure and content of a university World Wide Web site." Journal of Information Science, Vol. 25 No. 3, pp. 219-227.

Moed, H. F. (2002) The impact-factors debate: the ISI's uses and limits, Nature, Vol. 415, pp. 731-732.

Rodríguez Gairín, J. M. (1997). "Valorando el impacto de la informacion en Internet: AltaVista, el "Citation Index" de la Red," Revista Espanola de Documentacion Cientifica, Vol 20. pp. 175-181. Available: http://www.kronosdoc.com/publicacions/altavis.htm

Ronayne, C. (1997). "Research and the new universities: Towards mode 2."
ATSE Focus, 98. Available: http://www.atse.org.au/publications/focus/focus-ronayne.htm

Rousseau, R., (1997). "Sitations: an exploratory study," Cybermetrics, Vol. 1. Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Smith, A. and Thelwall, M. (2002). "Web Impact Factors for Australasian universities", Scientometrics, Vol. 54 No. 3, pp. 363-380.

Tang, R. and Thelwall, M. (2003). "Disciplinary differences in US academic departmental Web site interlinking", Library and Information Science Research, to appear.

Thelwall, M. and Tang, R. (2003). "Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan", Scientometrics, to appear.

Thelwall, M., & Vaughan, L. (2004, to appear), "A fair history of the Web? Examining country balance in the Internet Archive", Library & Information Science Research.

Thelwall, M. (2000). "Web Impact Factors and search engine coverage", Journal of Documentation, Vol. 56 No. 2, pp. 185-189.

Thelwall, M. (2001a). "A Web crawler design for data mining," Journal of Information Science, Vol. 27 No. 5, pp. 319-325.

Thelwall, M. (2001b). "A publicly accessible database of UK university Web site links and a discussion of the need for human intervention in Web crawling." Available: http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf

Thelwall, M. (2001c). "Extracting macroscopic information from Web links", Journal of the American Society for Information Science and Technology, Vol. 52 No. 13, pp. 1157-1168.

Thelwall, M. (2002a). "Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university Web sites", Journal of the American Society for Information Science and Technology, Vol. 53 No. 12, pp. 995-1005.

Thelwall, M. (2002b). "The top 100 linked pages on UK university Web sites: high inlink counts are not usually directly associated with quality scholarly content", Journal of Information Science, Vol. 28 No. 6, pp. 485-493.

UNESCO. (1997). "International Standard Classification of Education (ISCED 1997)." Available: http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm

Vaughan, L. and Thelwall, M. (2003). "Scholarly use of the Web: What are the key inducers of links to journal Web sites?" Journal of the American Society for Information Science and Technology, Vol. 54 No. 1, pp. 29-38.

Weare, C., and Lin, W. Y. (2000). "Content analysis of the World Wide Web-Opportunities and challenges." Social Science Computer Review, Vol. 18 No. 3, pp. 272-292.

Wilkinson, D., Harries, G., Thelwall, M. and Price, E. (2003). "Motivations for academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication", Journal of Information Science, Vol. 29 No. 1, pp. 59-66.

Appendix 1. Classifier instructions.

Each country will be classified independently at the same time by two classifiers, using the spreadsheet list. The classification should be continued *until 100 subject classifications have been made*. This will probably mean visiting 200-300 sites. In each group, some obvious classifications have already been made, mainly the university home domains, so these do not need to be revisited.

1. Each classifier will classify the domain according to one of the categories below. The classification should be based first on the home page description and if the content of the domain is not clear from this, by visiting other pages until it is. Note (1): most of the domains will need to have www. added to the start before being accessed. E.g. csu.edu.au should be changed to www.csu.edu.au before being accessed. For crawling purposes it has been assumed that sites with and without a starting www. are identical. Note (2): sites that have gone can often be visited using the Internet Archive's Wayback machine at www.archive.org, and should be classified in this way, if possible.
2. Each domain falling into the "Subject-based internally created site" category (expected to be the main one, e.g. departments, research groups.) will then be classified by subject according to Table 1 (below). Other domains (e.g. home sites) should be ignored for this stage. More information about the UNESCO codes can be gained from the more detailed classification scheme which will be sent as a Word document, or from the URL http://www.usc.es/citt/privado/codigos/unesco.htm (Spanish). Unknown subjects should be recorded as 0. Each domain should be given one main coding. ***If*** there is clearly related content from a different subject, then this should be given as a secondary coding in the next column. If there are further related codings then these should be listed in the third column. The objective of the classification is to *predict* what the content of the site is from the description on the home page. In some cases you may wish to visit other pages to clarify the content. This is not expected to be a 100% accurate process.
3. Each domain should be given a short one-sentence description in English that will enable the third classifier to arbitrate over disagreements in content. For non-classified domains, this should describe the domain so that it can be discussed in the paper. This could be a few words or a sentence.

Appendix 2. The Classification Scheme

Table 1. The revised (domain-based) Web site classification scheme and its relationship to the UNESCO categories

ID	Classification	Corresponding UNESCO categories
0	Unknown	None
1	Mathematics and logic	11 Logic; 12 Mathematics
2	Computing	1203 Computing
3	Physics and astronomy	21 Astronomy and astrophysics; 22 Physics
4	Chemistry	23 Chemistry
5	Life sciences	24 Life sciences
6	Earth and Space Sciences	25 Earth and Space Sciences
7	Agricultural sciences	31 Agricultural sciences
8	Medical sciences	32 Medical sciences
9	Technology and Engineering	33 Technology and Engineering
10	Anthropology	51 Anthropology
11	Economics	53 Economics
12	Geography & demography	54 Geography; 52 Demography
13	History	55 History
14	Law	56 Law
15	Linguistics	57 Linguistics
16	Education	58 Education
17	Political science	59 Political science
18	Psychology	61 Psychology
19	Arts	62 Arts
20	Sociology	63 Sociology
21	Philosophy and Ethics	71 Ethics